Estudo sobre o K-Means e uso do algoritmo como solução do mau descarte de lixo em Minas Gerais

Arthur Rodrigues Soares Quadros[1], **João Pedro Torres**[1], **Sarah Luiza de Souza Magalhães**[1]

[1] Pontifícia Universidade Católica de Minas Gerais

{aquadros,joao.silva,sarah.magalhaes.1280966}@sga.pucminas.br

Resumo. *Proposta de heurística customizada para solução de problemas reais de clusterização utilizando o algoritmo de Inteligência Artificial para aprendizado não supervisionado K-Means. Para isso, foi criado um algoritmo de clusterização capaz de agrupar dados bidimensionais, que possui a mesma proposta do K-Means, porém, com algumas mudanças em relação ao modo de seleção de centróides iniciais, cálculo de distância entre pontos e definição de convergência com uso de um percentual de tolerância. Os resultados finais, dentro do estudo de caso proposto, se mostram satisfatórios dada a simplicidade da heurística em comparação à implementada na biblioteca Python Scikit-learn. A Soma do Quadrado dos Erros (SSE) dessa heurística fica maior no total, porém, o erro por cluster, em alguns casos, fica muito menor.*

Palavras-chave: *K-Means; Clusterização; Heurísticas; Distância de Manhattan.*

1. Introdução

Dentre os problemas ambientais mais expressivos no cenário nacional, destaca-se a exorbitante quantidade de lixo eletrônico descartada incorretamente [de Andrade and Ferreira 2011]. Diante desse cenário, a implementação de unidades de tratamento mostra-se como uma solução viável para o problema. Contudo, restringindo o contexto de ambientação da proposta ao estado de Minas Gerais, nota-se a necessidade de mapeamento das regiões e municípios mais carentes dos serviços de tratamento especial de lixo eletrônico, haja vista que a quantidade e o posicionamento dessas estações não é conhecido. A partir dessa conjuntura, a proposta deste artigo, consistente no uso de algoritmos de Inteligência Artificial e estratégias de Aprendizado de Máquina não-supervisionado, visa realizar o agrupamento dos municípios mineiros em regiões de possível instalação de estações de tratamento especial de lixo.

2. Metodologia

Em relação às metodologias, há duas questões a serem analisadas: a primeira, de cunho computacional, sendo a análise da heurística implementada sobre o algoritmo *K-Means*, mas sem aplicabilidade na solução do problema em questão; e a segunda, a aplicabilidade do *K-Means* para solução do problema, com análise de custos totais e viabilidade da solução.

2.1. A herística

A implementação da heurística, metodologia para solução de um problema [Romanycia and Pelletier 1985], foi feita na linguagem de programação *Python*, tendo

sido escolhida por sua simplicidade de uso e abundância de recursos já nativos ou de rápida implementação. Todo o código, testes e visualização de resultados foram feitos através de suas bibliotecas externas, notavelmente: *Scikit-learn, Numpy, Pandas* e *Matplotlib*.

A proposta da heurística surge com a ideia de simplificar a execução do algoritmo *K-Means* com resultados aceitáveis para o contexto do problema em que, principalmente, não haja uma perda de representatividade em relação ao revés, nem à solução do algoritmo padrão usado para comparações. O algoritmo base utilizado para a avaliação de desempenho da heurística é o proposto na biblioteca *Scikit-learn*.

A heurística consiste em alguns procedimentos em diferentes etapas da implementação padrão do *K-Means*, nesse caso, sendo em grande parte baseada no cálculo da Distância de Manhattan entre os pontos (x, y), $(longitude, latitude)$, por ser simples de aumentar o número de dimenções, ideal para clusterização [Pandit et al. 2011]. Por isso, ele não deixa de ser uma variação desse algoritmo através de algumas abordagens matemáticas simplificadas. Essas etapas são: definição dos centróides iniciais, encontrar valores iniciais aproximados dos valores ótimos, cálculo da distância propriamente dita para seleção dos pontos para cada cluster, utilizando a Distância de Manhattan para potenciais agrupamentos mais do que bidimensionais, e a condição de parada utilizando uma métrica por tolerância, ao invés de uma aplicação da Norma de Frobenius ou Norma Matricial.

Na seleção dos centróides iniciais, essa heurística seleciona N grupos de centróides diferentes com posições aleatórias, isto é, seleciona K (número de clusters) pontos P pertencentes ao conjunto de pontos X contido nos dados, para depois montar N matrizes de afinidade simplificadas (utilizando o cálculo de Distância de Manhattan ao invés do Kernel de Gauss) entre todos eles, uma para cada grupo de centróides selecionados aleatoriamente. Após esse cálculo, temos a métrica de escolha de qual dos N grupos serão utilizados para início do cálculo, para isso, em cada uma das matrizes $K \times K$ geradas, calculamos a distância média entre todos os pontos em cada uma das matrizes, selecionando o grupo que possui a maior distância média de todas para serem os centróides iniciais. Essa ideia segue a premissa de que, quanto maior a distância entre centróides, melhor a área coberta por cada centróide tende a ser distribuída.

Após isso, com o grupo de centróides ideal escolhido, nós realizamos a primeira iteração para gerar os K *clusters* e, com eles definidos, nós recalculamos os centróides com uma média aritmética simples entre todas as coordenadas (x, y) pertencentes a cada *cluster*. Na primeira iteração, o teste para condição de parada não é realizado, mas, após atualizar os centróides uma primeira vez e iterar uma segunda vez, os valores pertencentes a cada *cluster* são atualizados e comparados com os resultados anteriores. Nesse momento, o teste de convergência é realizado. A tolerância T é um percentual que determina o quanto os valores precisam mudar para ser considerado uma mudança significativa, sendo que uma tolerância muito alta permite que dados com valores muito distantes ainda sejam considerados IGUAIS, enquanto uma tolerância muito baixa considerará que dados muito próximos ainda sejam considerados DIFERENTES. Seja v_o a distância entre um dado *cluster* e um ponto x qualquer da iteração antiga e v_n, de forma análoga, da iteração nova, para v_n ser considerado igual, ele precisa estar dentro de um intervalo fechado $[v_o(1-T), v_o(1+T)]$. Caso mais de 75% dos dados atualizados sejam consider-

ados iguais aos da iteração anterior de acordo com a tolerância determinada, uma parada antecipada é chamada, o que faz o algoritmo parar a execução mesmo se ainda não houver alcançado o número de iterações máximo determinado. Em resumo, esse algoritmo com a tolerância baixa fará mais iterações, com a tendência de otimizar mais o resultado final, enquanto, com uma tolerância mais alta, ele tenderá a acabar com a parada antecipada mais cedo. Em outras palavras, a tolerância mais alta determina convergência mais rápido; a mais baixa, mais lento.

2.2. Proposta de solução

Dado o grande volume de lixo produzido pelo estado de Minas Gerais, é nítida a necessidade de uma proposta que solucione esse problema, sem que faça o estado se submeter ao uso excessivo de recursos como lixões a céu aberto. Com isso, temos os dados utilizados para a solução do problema: informações de localização geográfica das cidades do estado [S. do Prado 2022], aliadas ao tamanho da população de cada uma de suas cidades, de acordo com as informações do censo de 2010 do IBGE [SUAS 2010].

2.2.1. Dados utilizados

Após todos os filtros nas duas bases de dados, são utilizadas três informações: longitude, latitude e porte (tamanho) de cada cidade do estado de Minas Gerais. Longitude e latitude, para montagem do mapa do estado e o porte para determinação dos pesos que cada cidade terá. Com o uso do porte das cidades, temos um fator extra que faz o algoritmo ter uma maior concentração em torno dos pontos $(longitude, latitude)$ das cidades de maior porte. Isso significa que, apesar de na implementação final do algoritmo e na representação gráfica do mapa, haver apenas duas dimensões, existem, na verdade três dimensões, com os hiperplanos $(longitude, latitude, porte)$, sendo que, originalmente, haviam cinco valores para o porte ("Pequeno I", "Pequeno II", "Médio", "Grande" e "Metrópole"), mas esses dados foram transformados em pesos para manipulação matemática, com o mapeamento dos valores transformando-os em (30, 30, 300, 700, 2800), ou seja, quanto maior o porte, maior o peso.

2.2.2. Parâmetros e erros do algoritmo

Em relação aos testes associados à heurística implementada, o número de *clusters* foi aumentado em 5, havendo, no total, 15 centróides. O erro medido para esses *clusters* não foi considerado fortemente para análise final dos dados, visto que isso conta com uma terceira dimensão inexistente para o cálculo das distâncias entre os pontos. No lugar disso, o que foi considerado para a análise final de qualidade de resultado foi a análise de viabilidade financeira, mostrada posteriormente, e a distribuição dos pontos considerando suas proximidades com os grandes centros urbanos, isto é, mais centróides (aterros sanitários) em torno de regiões com maiores populações.

2.2.3. Necessidade dos aterros em Minas Gerais

Não é novidade que os aterros sanitários são considerados uma das melhores opções para o descarte de Resíduos Sólidos Urbanos (RSU), dado que evitam grande parte dos problemas gerados pelos lixões a céu aberto, como a contaminação do ar, solo e das águas [Conde et al. 2014] na região em que se encontra. Não à toa, tornou-se lei (lei 12.305/2010, Política Nacional dos Resíduos Sólidos, PNRS) que o país deveria aderir ao uso de aterros sanitários como alternativa mais ambientalmente sustentável. Dada a implementação dessa lei no ano de 2010, essa transformação do tratamento de RSU no país vem acontecendo de forma gradual.

No caso de Minas Gerais, supondo que o estado siga a mesma proporção da região em que se encontra, temos que, em 2021, 73,4% do RSU gerado é descartado de forma adequada (através, principalmente, dos aterros sanitários), e o complementar, 26,6%, de forma inadequada (através, principalmente, dos lixões) [DOS RESÍDUOS 2021]. Isso significa que, de todo o lixo gerado no estado de Minas Gerais, mais de 25% ainda não está sendo descartado de acordo com a PNRS, contaminando o meio ambiente de forma desnecessariamente ampliada. Com isso, é plausível a consideração de uma redistribuição do descarte de lixo do estado, passando-se a incluir as cidades que não estão de acordo com a PNRS, além de reforçar o descarte adequado de lixo em maiores volumes nas cidades mais populosas.

2.2.4. Visualização da proposta

A proposta gerada pela aplicação do algoritmo utilizado como base nos testes da heurística, como explicado anteriormente, usa uma terceira dimensão inexistente e, por isso, a distribuição dos centróides, considerando somente o mapa bidimensional (sem o desenho do peso do porte da cidade), fica desproporcional. Mas, analisando a ditribuição dos centróides em relação às regiões mais populosas, é possível verificar que a região de conurbação da capital Belo Horizonte (a área com a maior concentração populacional do estado) possui uma atenção maior dos aterros.

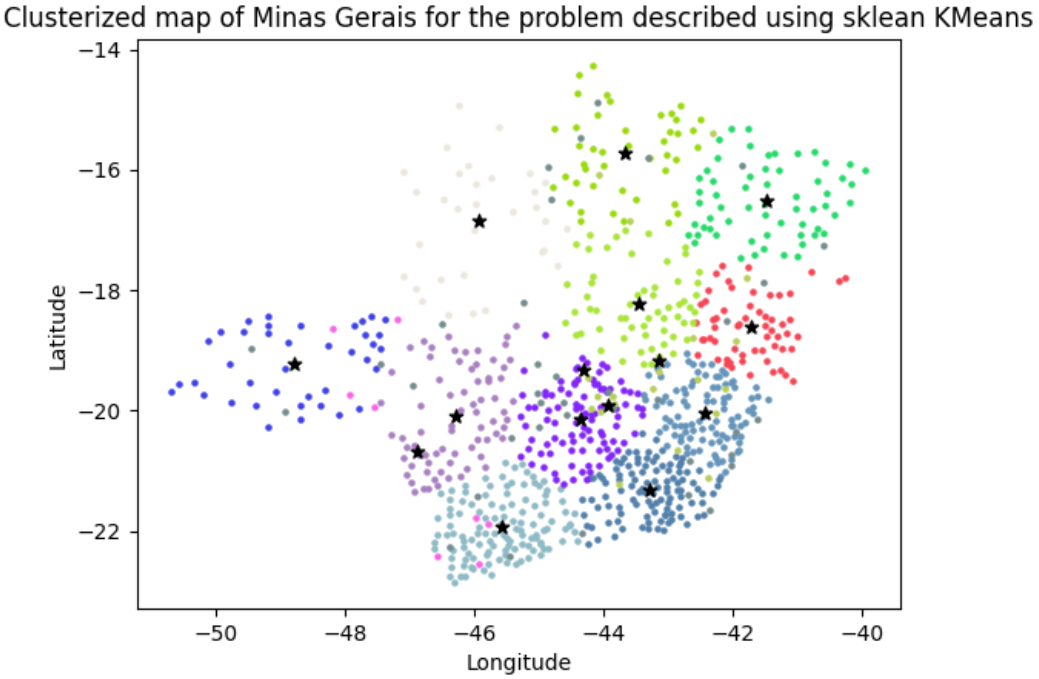

Figure 1. Distribuição considerando o porte das cidades

Nesse momento, é importante notar que, de acordo com o Centro Mineiro de Referência em Resíduos, já existem 17 aterros sanitários em atividade no estado, sendo que esses aterros atendem cerca de 50% das cidades de Minas Gerais. Por isso, a proposta inclui acrescentar a esses aterros sanitários os novos 15 em regiões próximas das coordenadas definidas pelos centróides da execução do algoritmo, em concordância com todas as definições de criação de aterros sanitários, como determinação matemática dos tamanhos, capacidades, projeções de validade e etc. determinados em [Spinola et al. 2017].

Após essa criação, é possível garantir que o restante das cidades no estado que ainda não estão conforme a PNRS passem a colaborar, fazendo com que o estado de Minas Gerais seja o mais de acordo com as políticas de coleta e tratamento de resíduos no país. Além de que tal solução possibilitaria uma interrupção do uso dos lixões a céu aberto no estado, com validade de décadas para frente.

3. Resultados

Para a análise dos resultados, é necessário separar o projeto em dois contextos: a avaliação da performance da heurística personalizada do *K-Means* em relação ao algoritmo base e a proposta de solução do problema de descarte de lixo no estado.

3.1. Sobre a heurística

Para a execução dos modelos, foram contemplados cinco hiper-parâmetros, atuantes, cada um, nas áreas de: aleatoriedade para seleção dos centróides iniciais (A); quantidade de grupos geográficos (*clusters*) a serem gerados (B); máximo de iterações com redefinição dos *clusters* e recálculo dos centróides (C); na verificação de desempenho para avaliação da parada antecipada (D) e na quantidade máxima de iterações (E). Partindo da utilização desses recursos, ambos os algoritmos foram executados de forma equiparável, de modo

que os resultados obtidos por ambos pudessem ser comparados de forma justa. Tais hiper-parâmetros podem ser visualizados na seguinte tabela.

nome	**valor**	**atuação**
n_clusters	10	B
n_init	50	A
max_iter	5000	C, E
tol	$1 \cdot 10^{-5}$	D, E
random_state	5	A

Table 1. Hiper-parâmetros e suas funções

Após executados os modelos, duas notórias diferenças foram observadas. A primeira, de cunho analítico, que pode ser facilmente notada a partir dos gráficos seguintes, diz respeito à diferença de centralidade dos centróides resultantes em relação aos seus respectivos *clusters*. A segunda, de cunho matemático, trata da diferença entre os valores J, dados pela expressão $\sum_{c=1}^{k} \sum_{x_j \in C_c} d(x_j,\ \bar{x}_c)^2$, de cada um dos modelos, que denotam a SSE calculada de cada um deles.

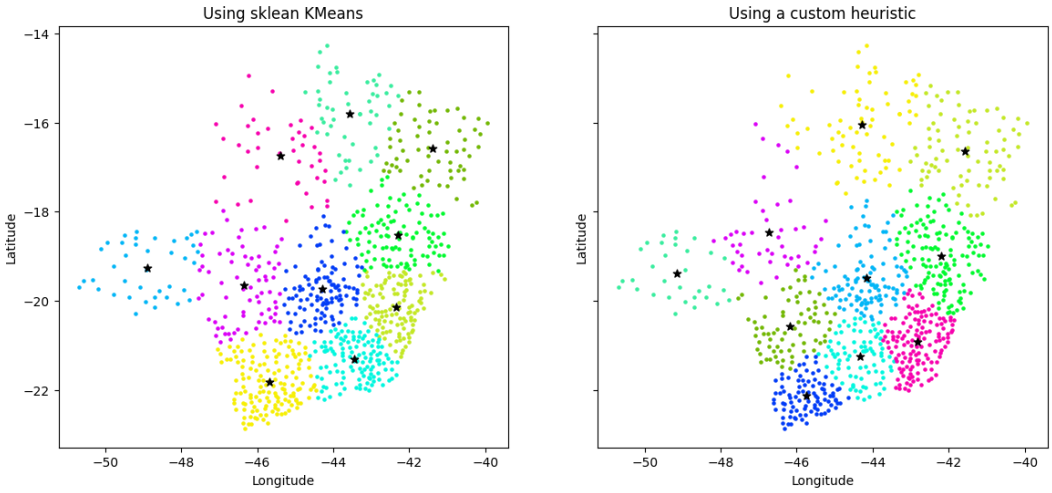

Figure 2. Comparação dos resultados

Ao avaliar essa diferença, nota-se que o modelo com essa heurística customizada gerou um J equivalente a aproximadamente 103,35% do SSE calculado pelo modelo base (647,49 em relação à base, 626,48), o que mostra que o primeiro conseguiu, com uma heurística consideravelmente mais simples, um resultado matematicamente muito próximo da base, considerando-se a métrica de erro total. Além disso, há também o SSE calculado para cada *cluster* individualmente, em que o erro individual médio no algoritmo base tende a ser melhor distribuído, com distâncias que variam entre 43,2 e 96,6, e valores intermediários muito abaixo da média entre os dois extremos[1], enquanto, nessa heurística, as distâncias variam entre 28,8 e 102,4, e valores intermediários levemente

[1]Todos os 10 valores do algoritmo base (com arredondamento para duas casas decimais), em ordem crescente de *clusters* k, $0 \leq\ k \leq\ 9$: 96,63, 50,59, 60,40, 71,95, 50,25, 60,97, 43,15, 63,58, 73,61 e 55,36.

abaixo da média entre os dois extremos[2], o que significa que, apesar de que alguns *clusters* possuem um valor consideravelmente menor do que o mínimo do algoritmo base, há outras distâncias que aumentam o erro total a ponto de ficar maior do que a base. Essa heurística consegue diminuir o erro individual de alguns *clusters* muito bem, mas à custa do erro individual de outros serem aumentados.

3.2. Proposta de solução

Com a proposta de solução gerada, caso seja tudo seguido de acordo, é possível que o descarte inadequado de lixo no estado seja completamente extinguído, além de que possuiría solução por décadas a partir do fim da implementação da proposta. Além disso, é também possível o uso do lixo acumulado nos aterros para geração de energia para o estado, o que significa, em última análise, uma fonte alternativa de energia que pode ser utilizada para baratear o custo de energia elétrica para a população do estado de Minas Gerais.

Visualmente, os resultados da implementação do algoritmo são satisfatórios, dada a distribuição através da terceira dimensão que são os portes das cidades.

Apesar de tudo isso, é importante lembrar que o preço de criação e manutenção de todos os aterros sanitários seria muito caro, principalmente sua criação, sendo estimado, com base nas estatísticas do projeto de [Spinola et al. 2017], um custo na casa das dezenas de bilhões de reais.

[2]Todos os 10 valores do algoritmo com a heurística customizada (com arredondamento para duas casas decimais), em ordem crescente de *clusters* k, $0 \leq k \leq 9$: 102,42, 94,81, 49,47, 102,18, 28,79, 38,93, 71,96, 33,73, 56,52 e 68,69.

References

Conde, T. T., Stachiw, R., and Ferreia, E. (2014). Aterro sanitário como alternativa para a preservação ambiental. *Revista Brasileira de Ciências da Amazônia/Brazilian Journal of Science of the Amazon*, 3(1):69–80.

de Andrade, R. M. and Ferreira, J. A. (2011). A gestão de resíduos sólidos urbanos no brasil frente às questões da globalização. *Rede-Revista Eletrônica do PRODEMA*, 6(1).

DOS RESÍDUOS, S. N. B. (2021). Panorama. *Associação Brasileira de Empresas de Limpeza Pública e Resíduos Especiais*.

Pandit, S., Gupta, S., et al. (2011). A comparative study on distance measuring approaches for clustering. *International journal of research in computer science*, 2(1):29–31.

Romanycia, M. H. and Pelletier, F. J. (1985). What is a heuristic? *Computational intelligence*, 1(1):47–58.

S. do Prado, K. (2022). Municípios brasileiros. Disponível em: `https://github.com/kelvins/Municipios-Brasileiros`. Acessado em: 20 de maio de 2023.

Spinola, G. M. R., de Andrade, P. R., and Nascimento, V. F. (2017). Caracterização e dimensionamento de aterros sanitários para resíduos sólidos urbanos no brasil e nos municípios paulistas. *Relatório final de projeto de iniciação científica. Inpe: São José dos Campos, SP*.

SUAS, R. (2010). Lista de municípios brasileiros e informações adicionais. Disponível em: `http://blog.mds.gov.br/redesuas/lista-de-municipios-brasileiros/`. Acessado em: 20 de maio de 2023.